SHAPE-MATCHED EVALUATION OF RETURNED SEGO CODING TREES ON MOLECULAR SCAFFOLD SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recognizing molecules with unfamiliar core frameworks is an important task for molecular property models. Hierarchical graph representations organize vertices into nested groups. Their detection performance, however, leaves open what the chosen groups reveal beyond the hierarchy's shape. We introduce a comparison that preserves group sizes and branching while randomizing vertex membership. We apply it to coding trees from a structural entropy guided graph detector and evaluate the edges retained within groups before representation learning. On molecular scaffold splits, these summaries yield positive average ranking gains over the random reference. The evidence depends on the detector and on how the reference is estimated: repeated evaluations retain positive means at the original randomization budget, while larger ensembles leave the gain uncertain. This comparison helps distinguish information in the chosen groups from that captured by a hierarchy of the same shape. The core code is provided in the supplementary material.

1 INTRODUCTION

Molecular property models encounter compounds whose core frameworks differ from those represented in training. Scaffold splits capture this challenge by separating molecules according to their ring systems and linkers (Bemis & Murcko, 1996; Wu et al., 2018; Hu et al., 2020). Graph out-of-distribution (OOD) detection seeks to distinguish unfamiliar molecules from the in-distribution (ID) population, raising the question of which structural features reveal these departures.

Recent methods use graph structure to identify unfamiliar samples. GOODAT extracts informative subgraphs at test time (Wang et al., 2024), while SEGO constructs coding trees through structural entropy optimization for multigrained contrastive learning (Hou et al., 2025a). These trees organize vertices into nested groups, specifying both their branching and membership. Overall detection performance leaves unresolved how much information the specific groupings provide beyond the hierarchy's shape.

Hierarchies with the same shape can place different vertices together. One grouping may retain many edges inside its groups, while another cuts across them. To evaluate this difference, we keep group sizes and branching fixed and randomize vertex membership. We then compare the edges retained by the returned coding tree with those retained under random assignments. This gives each tree a reference with the same shape.

SEGO's tree construction provides a concrete setting for this question: it selects groups by structural entropy before learning graph views. We evaluate edge summaries of these groups on molecular scaffold splits from the Open Graph Benchmark (OGB), using connectivity without atom or bond attributes. Matching by node count and pairing by mean degree reduces coarse structural differences. Separately fitted detectors rank the returned and randomized summaries on the same molecular pairs.

Our contribution is a shape-matched evaluation of the ranking information in hierarchy summaries. It distinguishes a construction's gain over random membership from its advantage over another construction, giving each its own reference. SEGO's mean gains are positive, while the gains of the two constructions do not clearly differ. Repeated pipelines retain positive means at a fixed randomization budget; larger budgets leave the gains uncertain. This evaluation makes the reference part of interpreting what a hierarchy's chosen groups reveal.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

Structural entropy describes a graph through nested groups of vertices (Li & Pan, 2016). This organization has been used to enhance graph views for node classification (Duan et al., 2024). SEGO brings coding trees into graph OOD detection, using a tree obtained by structural entropy optimization to guide contrastive learning at local, global, and tree levels (Hou et al., 2025a). Its detection results reflect the combined effects of the grouping, learned views, and scoring procedure. We examine the edges retained within the returned groups, using random assignments with the same hierarchy shape as a reference.

The relevant notion of an unfamiliar graph depends on the shift being studied. Graph anomaly detection distinguishes structural and attributed departures (Akoglu et al., 2015). GOOD and DrugOOD organize graph and molecular shifts into evaluation settings (Gui et al., 2022; Ji et al., 2023), while GOODAT addresses test-time detection by learning informative subgraphs from test samples (Wang et al., 2024). We study unfamiliar molecular core frameworks using OGB's scaffold splits (Hu et al., 2020). Omitting atom and bond attributes focuses the evaluation on connectivity.

A random reference is informative only in relation to the properties it preserves. Constrained network ensembles compare observed organization with variation allowed under fixed graph properties, such as a degree sequence (Maslov & Sneppen, 2002; Milo et al., 2002; Fosdick et al., 2018). Here the graph stays unchanged; randomization acts on vertex membership in the hierarchy. Preserving group sizes and branching avoids mixing membership with shape in the reference. It also means that a second hierarchy construction needs its own reference, because its shape can differ from SEGO's.

3 Returned Trees and Shape-Matched Representations

For each graph, we summarize edges retained by the returned groups and by random groups of the same shape. Separately fitted anomaly detectors compare their rankings on shared molecular pairs. We first define corresponding edge-fraction coordinates, then the full feature vectors used in the primary comparison (Figure 1).

3.1 From molecular connectivity to edge summaries

Let $G = (V, E)$ be the binary undirected adjacency graph of an OGB molecule. Repeated directed bond records are collapsed into one undirected edge, and self-loops are removed. Atom attributes, bond types, and property labels are unused. Graphs with fewer than two vertices or no edges are ineligible.

SEGO searches for a hierarchy with low structural entropy. We run its official greedy construction at the requested height (Hou et al., 2025b). We call its output the *returned tree* and its nested vertex groups the *returned hierarchy*. The objective assigns each group a cost based on its boundary edges and its volume relative to its parent. For a non-root node α, write α^- for its parent and V_α for its vertex set. Its boundary has g_α edges, and its volume is $\mathrm{vol}(V_\alpha) = \sum_{v \in V_\alpha} \deg(v)$. The implementation's smoothed contribution is

$$h_\alpha = -\frac{g_\alpha}{\mathrm{vol}(V)} \log_2 \frac{\mathrm{vol}(V_\alpha) + 1}{\mathrm{vol}(V_{\alpha^-}) + 1}. \tag{1}$$

The objective sums h_α over non-root nodes. We evaluate the tree returned by this search, without assuming a global optimum.

To compare group membership across hierarchies, we count how much connectivity remains inside a group at each level. Levels run upward from singleton leaves at 0 to the root at requested height k. Excluding these two trivial partitions leaves $\Pi_1^{\mathrm{ret}}, \ldots, \Pi_{k-1}^{\mathrm{ret}}$. Writing $c_\ell(v)$ for the group containing vertex v at level ℓ, the within-group edge fraction is

$$w_\ell(G, \Pi) = \frac{1}{|E|} \sum_{(u,v) \in E} \mathbf{1}\{c_\ell(u) = c_\ell(v)\}, \qquad w(G, \Pi) = [w_1, \ldots, w_{k-1}]. \tag{2}$$

A larger coordinate means that more edges remain inside groups at that level. This compact vector retains edge fractions, rather than all memberships or SEGO's learned views.

3.2 A RANDOM REFERENCE WITH THE SAME SHAPE

Each random draw uniformly permutes vertices among the lowest-level groups, preserving every group size. Copying the upper child lists preserves the branching above those groups. Thus the graph and hierarchy shape stay fixed while membership changes. We summarize $R = 64$ draws by the mean and spread of their within-group fractions $w^{(r)}$:

$$\mu_R(G) = \frac{1}{R}\sum_{r=1}^{R} w^{(r)}(G), \qquad \sigma_R(G) = \left[\frac{1}{R}\sum_{r=1}^{R}\big(w^{(r)}(G) - \mu_R(G)\big)^2\right]^{1/2}. \tag{3}$$

These elementwise moments describe the edge retention expected from randomized membership for this graph and hierarchy shape.

For $n = |V|$ and group sizes $s_{\ell j}$ at level ℓ, the exact ensemble mean is

$$\mu_{\infty,\ell} = \mathbb{E}[w_\ell^{(r)}] = \frac{\sum_j s_{\ell j}(s_{\ell j} - 1)}{n(n-1)}. \tag{4}$$

Every edge has this same probability of having both endpoints in one group, so linearity of expectation gives the identity. Conditional on group sizes, the mean is independent of edge arrangement. The identity characterizes the mean's target; the reported experiments retain sampled μ_R and σ_R. Increasing R therefore changes the estimated reference moments used by both feature branches below, testing the sensitivity of the fitted representation comparison.

The simplest comparison gives one detector the returned fractions w^{ret} and another the random mean μ_R. Their coordinates measure edge retention at the same levels, so we call this the shared-coordinate comparison. Each representation receives its own standardizer and fitted detector.

The primary comparison extends these fractions with reference information. The full returned representation includes deviations from the random mean; the random representation includes the ensemble spread:

$$\begin{aligned} x_k^{\mathrm{ret}}(G) &= \big[w^{\mathrm{ret}}(G),\ w^{\mathrm{ret}}(G) - \mu_R(G)\big], \\ x_k^{\mathrm{random}}(G) &= \big[\mu_R(G),\ \sigma_R(G)\big]. \end{aligned} \tag{5}$$

Both vectors have the same dimension, but their second halves have different meanings: a signed residual versus a standard deviation. Each receives its own standardizer and detector fitted on ID molecules. The shared-coordinate comparison checks whether the full-vector result also appears without these extra coordinates.

3.3 ALTERNATIVE CONSTRUCTION AND DEGREE FEATURES

We also evaluate a connectivity-guided hierarchy, called the proxy. At $k = 3$, it starts at the unvisited vertex with greatest degree and traverses breadth-first. Unvisited neighbors are ordered by decreasing degree, with ties broken by increasing vertex index. Consecutive traversal positions form groups of arity three. Consecutive triples of groups form supergroups under one root. The proxy's own shape-matched ensemble supplies the reference moments in

$$x_3^{\mathrm{proxy}} = \big[w^{\mathrm{proxy}},\ w^{\mathrm{proxy}} - \mu_R^{\mathrm{proxy}}\big], \quad x_3^{\mathrm{proxy\text{-}random}} = \big[\mu_R^{\mathrm{proxy}},\ \sigma_R^{\mathrm{proxy}}\big]. \tag{6}$$

The moments use 64 randomizations of the proxy hierarchy. We also compare w^{proxy} with μ_R^{proxy} using the shared within-group coordinates. Each construction is consequently measured relative to the random assignments allowed by its own shape.

Degree statistics give a simpler connectivity baseline against which to judge the hierarchy's absolute discrimination. The vector records mean degree, degree spread, maximum degree, and the fraction of vertices with degree at least three:

$$d(G) = \left[\mathrm{mean}(\deg), \mathrm{sd}(\deg), \max(\deg), |V|^{-1}\sum_{v\in V}\mathbf{1}\{\deg(v) \geq 3\}\right]. \tag{7}$$

Combined representations concatenate $d(G)$ with a tree representation, then fit one standardizer and detector to the concatenated vector.

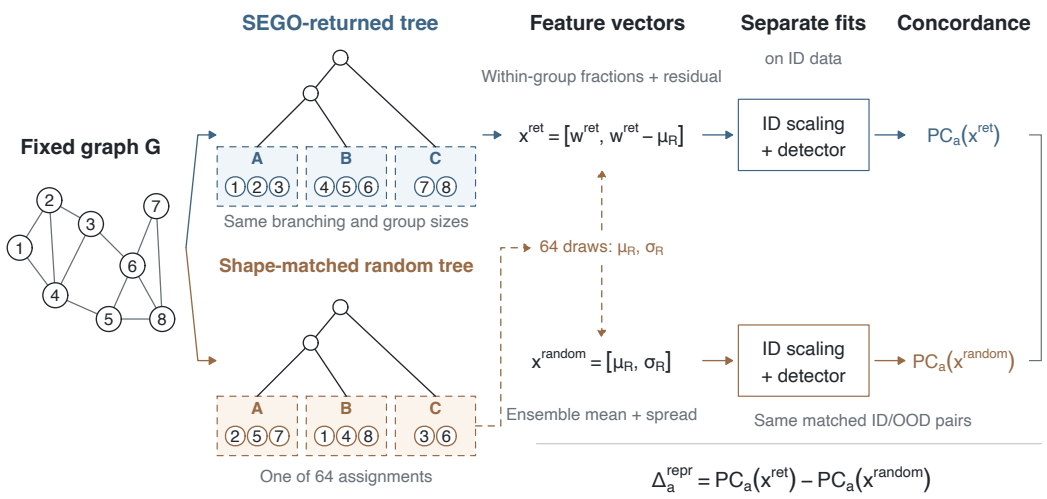


Figure 1: SEGO-returned tree and shape-matched random representations on an illustrative graph. SEGO supplies the upper coding tree; 64 random assignments preserve its branching and group sizes at the lowest level (3, 3, and 2) while changing vertex membership (one draw shown). Both arms use the same graph and matched ID/OOD pairs. Their feature vectors receive separate standardizers and detectors fitted on ID data before the paired concordance comparison.

4 Evaluation Design

One evaluation fits a detector on ID training molecules and asks it to rank an unfamiliar-scaffold molecule above a matched, held-out ID molecule. The returned and randomized representations use the same graph cohorts, fitting subset, and evaluation pairs; their feature values and fitted detectors differ. Repeating this comparison across detectors tests whether the ranking gain depends on how the summaries are scored.

4.1 Scaffold contexts and matched molecules

The five evaluation benchmarks are ogbg-molbace (1,513 eligible graphs), ogbg-molbbbp (2,039), ogbg-molsider (1,396 of 1,427), ogbg-moltox21 (7,812 of 7,831), and ogbg-molclintox (1,476 of 1,477). Official validation and test partitions define two OOD contexts per benchmark. The specification was developed on ogbg-moltoxcast (8,565 of 8,576), which is reported separately and excluded from the evaluation mean. Proxy-reference and feature-component analyses are subsequent sensitivity checks.

Training molecules supply disjoint detector-fitting and evaluation-ID pools. Within each exact node-count stratum s, a seed permutes the N_s eligible training molecules. The first $\max(1, \mathrm{round}(0.25N_s))$ enter the fitting subset; the rest are available for evaluation. We retain a stratum if it has at least three OOD molecules and enough evaluation-ID molecules to match them all.

Within each retained stratum, we process OOD molecules in seeded random order. Each is paired with the available ID molecule closest in mean degree, without replacement or a degree caliper. Seeds 17, 271, 911, 2027, and 8081 produce 11,955 pair occurrences across five benchmarks, two contexts, and five matchings.

Exact node-count matching removes graph size as a direct distinction within a pair; nearest-degree matching reduces a second coarse cue. It retains 2,391 of 2,860 OOD molecules and uses 4,500 unique evaluation-ID molecules. The mean absolute standardized difference in mean degree falls from 0.632 before matching to 0.091 afterwards, and the median paired degree gap is zero (Table 5). These diagnostics describe the population on which the hierarchy comparison is made.

4.2 Fitted detectors and ranking contrasts

All standardizers and detectors are fitted separately for each benchmark, context, seed, and representation. Each coordinate is centered by its fitting-ID mean and divided by its standard deviation plus 10^{-6}. Ledoit–Wolf scores squared distance using shrinkage covariance (Ledoit & Wolf, 2004). kNN-5 averages Euclidean distance to the five nearest fitting vectors (Cover & Hart, 1967; Ruff et al., 2021). Isolation Forest uses 200 trees and detector seed zero (Liu et al., 2008). These cover covariance-based, local-distance, and isolation-based scoring.

Paired concordance counts how often the OOD molecule receives the higher anomaly score, giving half credit to ties. For detector a, score $s_a(G; x)$ increases with anomalousness; over M matched pairs,

$$\mathrm{PC}_a(x) = \frac{1}{M} \sum_{m=1}^{M} \left[\mathbf{1}\{s_a(G_{\mathrm{ood},m}; x) > s_a(G_{\mathrm{id},m}; x)\} + \tfrac{1}{2}\mathbf{1}\{s_a(G_{\mathrm{ood},m}; x) = s_a(G_{\mathrm{id},m}; x)\} \right]. \tag{8}$$

Exchangeable paired scores have expected concordance 0.5. Subtracting concordances on the same pairs measures the ranking gain from the returned representation:

$$\Delta_a^{\mathrm{repr}} = \mathrm{PC}_a(x_k^{\mathrm{ret}}) - \mathrm{PC}_a(x_k^{\mathrm{random}}), \qquad \Delta_a^{\mathrm{within}} = \mathrm{PC}_a(w^{\mathrm{ret}}) - \mathrm{PC}_a(\mu_R). \tag{9}$$

The first contrast uses the full vectors; the second uses only matched within-group coordinates. Proxy contrasts substitute the corresponding proxy quantities. Because each term uses a separately fitted detector, these are comparisons of fitted representations, not causal effects of membership with a fixed scorer.

4.3 Aggregation and uncertainty

We average over the two contexts and five matchings within each benchmark, then weight benchmarks equally so that Tox21's larger collection does not dominate:

$$\mathrm{PC}_a = \frac{1}{5} \sum_{b=1}^{5} \frac{1}{10} \sum_{c=1}^{2} \sum_{q=1}^{5} \mathrm{PC}_{a,b,c,q}\,. \tag{10}$$

The primary setting is $k = 3$, $R = 64$, selected during ToxCast development. A height analysis constructs new trees at $k = 2$ and $k = 4$ for each detector; it changes the hierarchy as well as the number of edge-fraction coordinates.

An OOD molecule can appear in several matchings. The primary paired bootstrap therefore resamples OOD identities within benchmark and context, applying the same multiplicities to all representations and matching seeds. We report percentile intervals from 5,000 replicates (Efron & Tibshirani, 1993). Clustering on ID identity provides a second dependence analysis. These intervals condition on the realized trees, random ensembles, fits, and matchings. Below, crossed bootstraps additionally resample randomization batches or complete pipeline realizations. Intervals are reported per detector without multiplicity adjustment.

5 Results

5.1 Representation gains and construction comparisons

At $k = 3$, $R = 64$, the returned representation improves mean paired concordance over its shape-matched reference under all three detectors (Figure 2a and Table 1). Gains range from $+0.0187$ to $+0.0218$. Only the Ledoit–Wolf interval clustered by OOD identity excludes zero. The proxy also has positive point gains over its own reference, but all its intervals span zero.

Comparing the two constructions requires a further distinction. Their gains over their respective references do not clearly differ for any detector (Table 6). Directly subtracting their absolute concordances gives a different result: the interval excludes zero under Isolation Forest alone. ToxCast, reported separately, has positive returned-tree gains and mixed proxy signs.

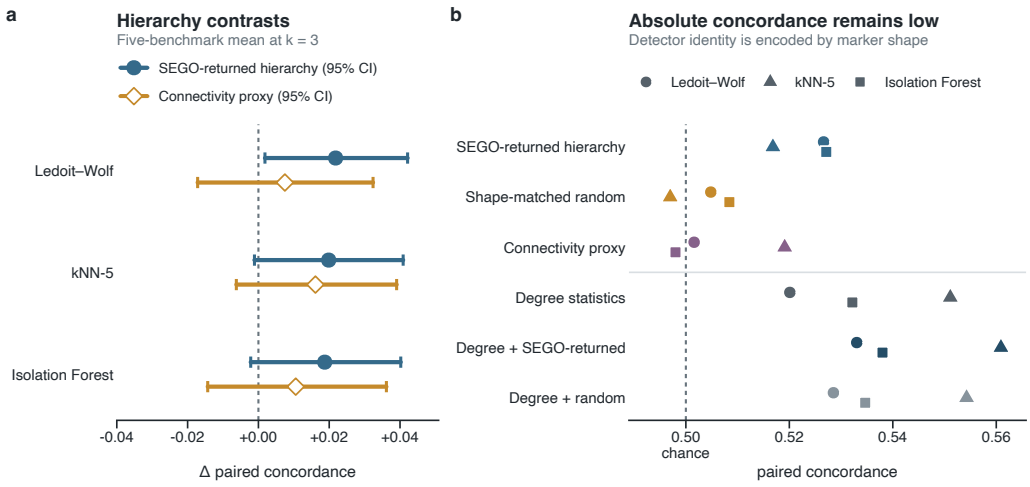


Figure 2: SEGO-returned hierarchy contrasts and absolute discrimination under the primary $k = 3$, $R = 64$ analysis. **(a)** Full SEGO-returned and proxy contrasts, each against the shape-matched random ensemble generated from the same hierarchy construction; bars are 95% paired intervals clustered by OOD identity. **(b)** Absolute paired concordance for six selected representations from Table 3; rows identify representations and marker shape identifies the detector. The SEGO-returned contrast is positive under all three detectors, but absolute discrimination remains near chance and degree statistics outperform it for kNN-5 and Isolation Forest. The means across five benchmarks give each benchmark equal weight; the analysis contains 11,955 matched pair occurrences across two contexts and five matching seeds.

Table 1: Concordance gains over each construction's own shape-matched random hierarchy under the primary $k = 3$, $R = 64$ analysis. Bracketed 95% intervals are paired bootstrap intervals clustered by OOD identity on the five evaluation benchmarks; ToxCast entries for the development benchmark are point estimates.

Benchmarks	Detector	SEGO-returned hierarchy	Connectivity-guided proxy
Evaluation mean	Ledoit–Wolf	$+0.0218\,[0.0018,\ 0.0421]$	$+0.0075\,[-0.0172,\ 0.0324]$
	kNN-5	$+0.0199\,[-0.0011,\ 0.0409]$	$+0.0161\,[-0.0062,\ 0.0390]$
	Isolation Forest	$+0.0187\,[-0.0022,\ 0.0402]$	$+0.0105\,[-0.0143,\ 0.0362]$
ToxCast	Ledoit–Wolf	$+0.0291$	-0.0242
	kNN-5	$+0.0354$	$+0.0049$
	Isolation Forest	$+0.0244$	-0.0191

The positive gain also appears when both representations use only the shared within-group coordinates. Gains range from $+0.0188$ to $+0.0236$, with only the Ledoit–Wolf interval above zero (Table 10). Adding the residual to the returned fractions gives no consistent gain across detectors. Comparing residuals alone with random spreads also leaves the gain uncertain. For the shared coordinates, the returned-minus-proxy gain differences range from $+0.0036$ to $+0.0108$; all intervals span zero (Table 6).

5.2 Pipeline reproducibility and reference estimation

Table 2 separates fixed-budget pipeline variation from sensitivity to reference estimation. Across 20 pipelines at $k = 3$, $R = 64$, we vary vertex labels, returned trees, ensembles, fitting subsets, matchings, and detector fits. Each context retains five matchings, and the official scaffold cohorts stay fixed. Mean gains range from $+0.0158$ to $+0.0274$ across detectors (Table 14). Intervals resampling both pipelines and OOD identities exclude zero for all three detectors, although kNN-5 has two nonpositive individual runs. The positive mean therefore persists across these perturbations at the original randomization budget.

Increasing R asks whether the result changes as the random reference is estimated more precisely. We use five independent streams per graph at $R = 128, 256, 512, 1024, 2048$. Within each stream, smaller ensembles are prefixes of larger ones. Both feature branches use the resulting moments. We assess precision using the between-batch standard deviation and the change in mean gain from the preceding R. The target, set before these runs, requires each quantity to be at most 25% of the mean OOD-cluster bootstrap standard deviation. A crossed bootstrap resamples both batches and OOD identities.

All five stream estimates remain positive at each tested R (Table 11). At $R = 2048$, mean gains range from $+0.0083$ to $+0.0128$, below the primary estimates, and all crossed 95% intervals span zero. Meanwhile, reference moments become more stable. Across graphs and coordinates, the root-mean-square between-batch standard deviation falls from 0.00822 to 0.00207 for μ_R and from 0.00550 to 0.00138 for σ_R. Only kNN-5 meets the detector-level precision target. Greater precision in the moments thus coexists with uncertainty in the ranking gain.

Table 2: SEGO-returned gains over the shape-matched reference at $k = 3$, using full representations. Cells give means and 95% intervals. The primary interval resamples OOD identities; the other columns jointly resample OOD identities and pipelines or randomization streams, respectively. All settings use the same official scaffold cohorts.

Detector	Primary $R = 64$	20 pipelines $R = 64$	5 streams $R = 2048$
Ledoit–Wolf	$+0.0218$ $[0.0018, 0.0421]$	$+0.0264$ $[0.0158, 0.0373]$	$+0.0097$ $[-0.0089, 0.0283]$
kNN-5	$+0.0199$ $[-0.0011, 0.0409]$	$+0.0158$ $[0.0049, 0.0271]$	$+0.0128$ $[-0.0053, 0.0311]$
Isolation Forest	$+0.0187$ $[-0.0022, 0.0402]$	$+0.0274$ $[0.0154, 0.0394]$	$+0.0083$ $[-0.0118, 0.0291]$

We next compare the returned tree with individual random trees, rather than their ensemble mean. Each of 256 draws from the first stream receives a separate standardizer and detector fitted to $w^{(r)}$. We compare these rankings with the fitted w^{ret} pipeline on the same pairs (Table 12). Mean gains are positive for all three detectors, and only the Ledoit–Wolf interval clustered by OOD identity excludes zero. None of the 256 random pipelines reaches the returned pipeline's concordance under Ledoit–Wolf; 12 do under kNN-5 and 5 under Isolation Forest. These counts describe the reference distribution. They are not randomization-test p values because the optimized hierarchy is not exchangeable with uniform assignments.

Vertex IDs provide a further source of variation even when molecular connectivity stays unchanged. We relabel each graph three times, rebuild both hierarchies, and map partitions back to the original IDs. We then refit the within-group-only detectors while keeping matching fixed. Across 42,708 graph–relabeling combinations, returned and proxy vectors remain unchanged in 45.9% and 46.6% of cases. The corresponding fractions for partitions at both informative levels are 16.7% and 15.9%. Raw returned-minus-proxy concordance differences stay positive in all three relabelings. Their magnitude changes: the kNN-5 difference rises from 0.0079 to 0.0270–0.0323 (Table 13).

5.3 Benchmark and tree-height variation

At $k = 3$, every evaluation benchmark has a positive point gain under kNN-5 and Isolation Forest (Figure 3). Ledoit–Wolf is positive on four of five; SIDER has a small negative estimate. ToxCast is shown separately because it informed the evaluation design. Table 9 gives all benchmark-level values.

The mean point gain decreases from $k = 2$ to $k = 4$ for every detector (Figure 4). At $k = 2$, gains range from $+0.0212$ to $+0.0358$, and all three intervals exclude zero. At $k = 4$, they range from $+0.0033$ to $+0.0113$, with all intervals spanning zero. Table 8 gives the full detector-by-height results. For Ledoit–Wolf, OOD-cluster, ID-cluster, and unpaired resampling agree on which heights have intervals above zero (Table 7). Each height uses a separately constructed tree; we do not directly test differences between heights.

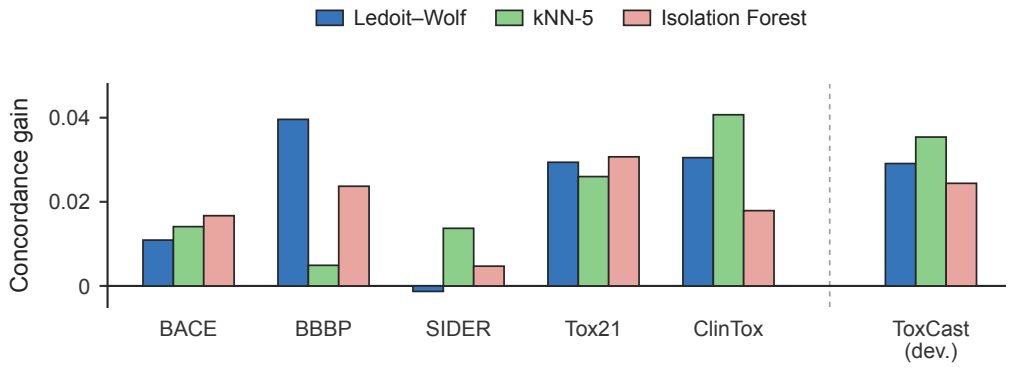


Figure 3: Paired concordance gains of the SEGO-returned hierarchy over shape-matched random for each benchmark at $k = 3$. Grouped bars show point estimates for each detector on a common scale centered at zero; the one negative estimate is on SIDER under Ledoit–Wolf. ToxCast is spaced apart as the development benchmark and is excluded from the evaluation mean across five benchmarks. Exact values are listed in Table 9.

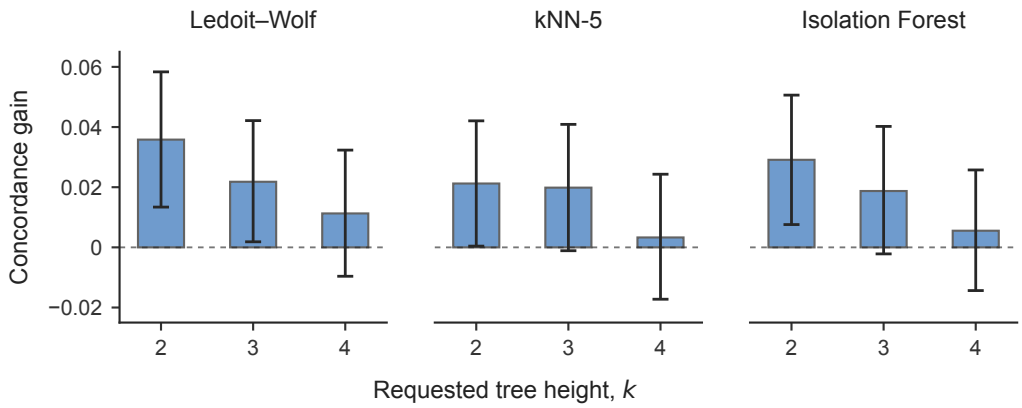


Figure 4: SEGO-returned hierarchy concordance gains over shape-matched random for separately constructed trees at $k = 2, 3, 4$. Panels share a common scale; bars are means across five benchmarks and whiskers are 95% paired intervals clustered by OOD identity. Each detector has its largest point estimate at $k = 2$; dashed horizontal lines mark zero.

5.4 Absolute discrimination and degree context

A positive gain over randomized membership can coexist with weak absolute detection. Degree statistics outperform the returned hierarchy alone under kNN-5 and Isolation Forest; the returned hierarchy leads narrowly under Ledoit–Wolf (Table 3). Matching reduces mean-degree differences, but the baseline also captures spread, maximum degree, and the fraction of higher-degree vertices. Degree plus returned features gives the largest mean concordance for each detector, reaching 0.5609 under kNN-5, still close to chance.

We also compare the two hierarchy representations after adding degree features to each. Subtracting degree-plus-random concordance from degree-plus-returned concordance gives a gain of 0.0045 for Ledoit–Wolf (95% interval $[-0.0119, 0.0212]$). The gain is 0.0066 for kNN-5 ($[-0.0103, 0.0237]$) and 0.0033 for Isolation Forest ($[-0.0156, 0.0227]$). All intervals span zero, leaving the gain over the random reference uncertain when both representations include degree information.

Table 3: Mean paired concordance of the fitted feature representations on the five evaluation benchmarks at $k = 3$; chance performance is 0.5.

Representation	Ledoit–Wolf	kNN-5	Isolation Forest
SEGO-returned edge features	0.5266	0.5168	0.5271
SEGO shape-matched summary	0.5048	0.4970	0.5084
Connectivity-guided edge features	0.5016	0.5191	0.4980
Proxy shape-matched summary	0.4941	0.5030	0.4875
Degree statistics	0.5201	0.5511	0.5322
Degree + SEGO-returned features	0.5330	0.5609	0.5380
Degree + SEGO random summary	0.5285	0.5543	0.5347

Synthetic graphs with known separation yield concordance 1.000 for all three detectors; random ID/OOD label swaps within matched pairs give approximately 0.5. These controls check ranking at clear separation and at chance. Table 4 reports entropy reproduction, tree construction, and coverage checks.

6 Discussion and Limitations

A construction's gain over randomized membership and its advantage over another construction are different comparisons. Shape matching gives each construction its own reference for the first comparison. Comparing those gains measures how much each improves over its own reference. In our evaluation, positive mean gains coexist with uncertain differences between constructions and near-chance absolute discrimination. The shared-coordinate, single-tree, and degree-feature comparisons clarify which part of this pattern depends on the representation and reference, before SEGO adds learned graph views.

The conclusions concern attribute-free edge summaries scored by separately fitted detectors, before SEGO's learned views. Matching restricts inference to the retained node-count strata. Pipeline repeats keep the official scaffold cohorts fixed. The proxy is tested only at $k = 3$, and ToxCast and individual-benchmark contrasts are reported as point estimates.

The primary intervals condition on one realized ensemble, while crossed intervals include batch or pipeline variation. At $R = 2048$, all gain intervals span zero and only kNN-5 meets the precision target. The clustering unit also matters: clustering by OOD identity leaves the primary kNN-5 and Isolation Forest gains uncertain, whereas clustering by ID identity gives intervals above zero (Table 6). Neither accounts explicitly for dependence within scaffold families. Intervals receive no multiplicity adjustment.

7 Conclusion

We present a shape-matched evaluation of the ranking information in hierarchy summaries. Applied to SEGO coding trees on molecular scaffold shifts, it distinguishes gains over randomized membership from differences between constructions. The evidence for positive gains depends on the detector and reference estimation, making both essential to interpreting what these summaries reveal.

AI Use Statement

Language model assistance was used primarily to improve the clarity, fluency, and organization of the writing.

References

Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Liang Duan, Xiang Chen, Wenjie Liu, Daliang Liu, Kun Yue, and Angsheng Li. Structural entropy based graph structure learning for node classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8372–8379, 2024.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Bailey K Fosdick, Daniel B Larremore, Joel Nishimura, and Johan Ugander. Configuring random graph models with fixed degree sequences. *SIAM Review*, 60(2):315–355, 2018.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Yue Hou, He Zhu, Ruomei Liu, Yingke Su, Jinxiang Xia, Junran Wu, and Ke Xu. Structural entropy guided unsupervised graph out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17258–17266, 2025a.

Yue Hou, He Zhu, Ruomei Liu, Yingke Su, Jinxiang Xia, Junran Wu, and Ke Xu. Official implementation of structural entropy guided unsupervised graph out-of-distribution detection, 2025b. URL `https://github.com/name-is-what/SEGO`. GitHub repository.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for machine learning on graphs. *Advances in Neural Information Processing Systems*, 33:22118–22133, 2020.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, Houtim Lai, Wei Liu, Junzhou Huang, Shuigeng Zhou, Ping Luo, Peilin Zhao, and Yatao Bian. DrugOOD: Out-of-distribution dataset curator and benchmark for AI-aided drug discovery—a focus on affinity prediction problems with noise annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8023–8031, 2023.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Angsheng Li and Yicheng Pan. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 62(6):3290–3339, 2016.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Luzhi Wang, Dongxiao He, He Zhang, Yixin Liu, Wenjie Wang, Shirui Pan, Di Jin, and Tat-Seng Chua. GOODAT: Towards test-time graph out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15537–15545, 2024.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

A Additional Results and Reproducibility Details

The following tables separate checks on the evaluation from comparisons of representations. Table 4 checks construction and ranking behavior, and Table 5 describes the retained molecules. In Table 6, D_{raw} subtracts the two constructions' absolute concordances; D_{gain} subtracts their gains over their own shape-matched ensembles. A difference between those gains is the relevant comparison when each construction has a different hierarchy shape.

Table 4: Validation checks for hierarchy construction and evaluation at $k = 3$.

Check	Observed result
Positive control with known separation	Concordance $= 1.000$ for all three detectors
Negative control: random ID/OOD label swaps within matched pairs	Concordance $= 0.5011$ (Ledoit–Wolf), 0.5006 (kNN-5), and 0.5003 (Isolation Forest)
Independent entropy calculation	Maximum absolute difference from the independent calculation: 1.8×10^{-15}
SEGO-returned tree entropy	Exceeded the mean from random assignments in $0/200$ audited graphs
Tree construction	No construction failures across all six benchmarks
Evaluation coverage	All 60 combinations of benchmark, context, and seed were retained, including the ToxCast development benchmark

Table 5: Coverage and balance of the graph matching procedure, aggregated over combinations of benchmark, context, and seed. Standardized differences compare mean degree between OOD and evaluation ID molecules within retained strata of node count.

Diagnostic	Observed result
Unique evaluation molecules	2,391 OOD; 4,500 ID
OOD retention	2,391 of 2,860 molecules (83.6%)
Retained strata of node count	243 of 438 strata (55.5%)
Absolute paired gap in mean degree	Mean 0.0097; median 0; 95th percentile 0.0741; maximum 0.5000
Mean absolute standardized difference	0.632 before matching; 0.091 after matching
Largest standardized difference after matching	0.269

Table 6: Hierarchy construction contrasts at $k = 3$. Point estimates are followed by 95% paired intervals clustered by OOD identity or ID identity. D_{raw} compares absolute concordances, whereas D_{gain} compares the two gains after referencing each construction to its own shape-matched random ensemble.

Contrast	Detector	Point	OOD clustered	ID clustered
SEGO-returned	Ledoit–Wolf	$+0.0218$	$[+0.0018, +0.0421]$	$[+0.0049, +0.0379]$
	kNN-5	$+0.0199$	$[-0.0011, +0.0409]$	$[+0.0017, +0.0385]$
	Isolation Forest	$+0.0187$	$[-0.0022, +0.0402]$	$[+0.0009, +0.0365]$
Connectivity-guided	Ledoit–Wolf	$+0.0075$	$[-0.0172, +0.0324]$	$[-0.0127, +0.0272]$
	kNN-5	$+0.0161$	$[-0.0062, +0.0390]$	$[-0.0033, +0.0354]$
	Isolation Forest	$+0.0105$	$[-0.0143, +0.0362]$	$[-0.0095, +0.0315]$
D_{raw}, full representation	Ledoit–Wolf	$+0.0250$	$[-0.0024, +0.0515]$	$[+0.0038, +0.0468]$
	kNN-5	-0.0023	$[-0.0255, +0.0213]$	$[-0.0215, +0.0166]$
	Isolation Forest	$+0.0291$	$[+0.0029, +0.0553]$	$[+0.0081, +0.0500]$
D_{gain}, full representation	Ledoit–Wolf	$+0.0143$	$[-0.0182, +0.0474]$	$[-0.0116, +0.0409]$
	kNN-5	$+0.0037$	$[-0.0276, +0.0375]$	$[-0.0231, +0.0305]$
	Isolation Forest	$+0.0082$	$[-0.0236, +0.0430]$	$[-0.0190, +0.0355]$
D_{gain}, within-group fractions	Ledoit–Wolf	$+0.0108$	$[-0.0238, +0.0461]$	$[-0.0175, +0.0390]$
	kNN-5	$+0.0045$	$[-0.0289, +0.0378]$	$[-0.0225, +0.0322]$
	Isolation Forest	$+0.0036$	$[-0.0295, +0.0380]$	$[-0.0227, +0.0312]$

Table 7: Ledoit–Wolf concordance gain of the SEGO-returned hierarchy over shape-matched random across tree heights. Brackets give 95% bootstrap intervals; the interval clustered by OOD identity is primary. The unpaired variant independently resamples clusters of OOD identities for the two representations.

k	Δ	OOD clustered	ID clustered	Unpaired
2	$+0.0358$	$[0.0134,\ 0.0583]$	$[0.0175,\ 0.0541]$	$[0.0144,\ 0.0577]$
3	$+0.0218$	$[0.0018,\ 0.0421]$	$[0.0049,\ 0.0379]$	$[0.0005,\ 0.0447]$
4	$+0.0113$	$[-0.0096,\ 0.0323]$	$[-0.0064,\ 0.0291]$	$[-0.0102,\ 0.0335]$

Tables 8 and 9 expand the height and benchmark comparisons shown in the main figures. Table 10 isolates the feature choices in the full representation: within-group fractions, their residuals from the random mean, and their concatenation. Every contrast uses separately fitted detectors evaluated on shared pairs.

Table 8: SEGO-returned hierarchy concordance gain over the shape-matched random hierarchy by tree height. Brackets give 95% paired intervals clustered by OOD identity.

Detector	$k = 2$	$k = 3$	$k = 4$
Ledoit–Wolf	$+0.0358$ $[0.0134, 0.0583]$	$+0.0218$ $[0.0018, 0.0421]$	$+0.0113$ $[-0.0096, 0.0323]$
kNN-5	$+0.0212$ $[0.0004, 0.0421]$	$+0.0199$ $[-0.0011, 0.0409]$	$+0.0033$ $[-0.0172, 0.0243]$
Isolation Forest	$+0.0291$ $[0.0076, 0.0506]$	$+0.0187$ $[-0.0022, 0.0402]$	$+0.0055$ $[-0.0144, 0.0257]$

Table 9: SEGO-returned hierarchy concordance gain over the shape-matched random hierarchy on each benchmark at $k = 3$. ToxCast is reported separately as the development benchmark.

Benchmark	Ledoit–Wolf	kNN-5	Isolation Forest
BACE	$+0.0109$	$+0.0141$	$+0.0167$
BBBP	$+0.0396$	$+0.0049$	$+0.0237$
SIDER	-0.0013	$+0.0137$	$+0.0047$
Tox21	$+0.0294$	$+0.0260$	$+0.0307$
ClinTox	$+0.0305$	$+0.0407$	$+0.0179$
ToxCast	$+0.0291$	$+0.0354$	$+0.0244$

Table 10: Mean concordance contrasts for feature component ablations at $k = 3$. The first two rows compare feature families of equal dimension; the last compares the full representation with its within-group fractions. Brackets give 95% paired intervals clustered by OOD identity.

Contrast	Ledoit–Wolf	kNN-5	Isolation Forest
Within: w^{ret} vs. μ_R	$+0.0236$ $[0.0016, 0.0453]$	$+0.0218$ $[-0.0013, 0.0442]$	$+0.0188$ $[-0.0029, 0.0405]$
Residual vs. σ_R	$+0.0229$ $[-0.0022, 0.0491]$	$+0.0087$ $[-0.0154, 0.0331]$	$+0.0197$ $[-0.0040, 0.0445]$
Full returned vs. w^{ret}	-0.0107 $[-0.0273, 0.0063]$	$+0.0057$ $[-0.0106, 0.0226]$	$+0.0082$ $[-0.0059, 0.0214]$

Reference and pipeline variations have different sampling units. Table 11 varies independently generated ensembles; Table 12 fits individual random trees. Table 13 changes vertex IDs while fixing matching, whereas Table 14 also varies ensembles, fitting subsets, matchings, and detector fits. The crossed intervals describe averages over the indicated batches or realizations on the fixed scaffold cohorts.

Table 11: Gain from the full representation by randomization count. The $R = 64$ row is the single stream specification selected during development; entries for $R \geq 128$ are means (standard deviations across batches) over five independent streams. The final row jointly resamples OOD identities and randomization batches.

R	Ledoit–Wolf	kNN-5	Isolation Forest
64 (primary)	$+0.0218$	$+0.0199$	$+0.0187$
128	$+0.0158\,(0.0050)$	$+0.0164\,(0.0057)$	$+0.0187\,(0.0044)$
256	$+0.0107\,(0.0072)$	$+0.0183\,(0.0033)$	$+0.0128\,(0.0047)$
512	$+0.0135\,(0.0050)$	$+0.0146\,(0.0034)$	$+0.0102\,(0.0030)$
1024	$+0.0091\,(0.0040)$	$+0.0144\,(0.0031)$	$+0.0128\,(0.0033)$
2048	$+0.0097\,(0.0064)$	$+0.0128\,(0.0006)$	$+0.0083\,(0.0044)$
2048 crossed 95% CI	$[-0.0089, 0.0283]$	$[-0.0053, 0.0311]$	$[-0.0118, 0.0291]$

Table 12: Individually fitted single-tree reference on within-group fractions at $k = 3$. The null range covers 95% of the concordances from 256 draws; the gain interval resamples OOD identities conditional on those draws. Exceedances are descriptive.

Detector	Returned	Null mean	Null 95% range	Exceed	Gain [OOD CI]
Ledoit–Wolf	0.5373	0.5045	$[0.4833, 0.5198]$	0	$+0.0328\,[0.0155, 0.0500]$
kNN-5	0.5112	0.4975	$[0.4816, 0.5142]$	12	$+0.0137\,[-0.0054, 0.0322]$
Isolation Forest	0.5190	0.5028	$[0.4869, 0.5183]$	5	$+0.0162\,[-0.0007, 0.0326]$

Table 13: Node ID relabeling sensitivity on within-group fractions at $k = 3$. Ranges cover three relabelings with separate fits and fixed matching. The raw difference is not a gain adjusted for hierarchy shape.

Detector	Returned PC: baseline → range	Proxy PC: baseline → range	Raw difference: baseline → range
Ledoit–Wolf	$0.5373 \to [0.5251, 0.5344]$	$0.4936 \to [0.4735, 0.4787]$	$0.0437 \to [0.0502, 0.0569]$
kNN-5	$0.5112 \to [0.5110, 0.5265]$	$0.5033 \to [0.4794, 0.4996]$	$0.0079 \to [0.0270, 0.0323]$
Isolation Forest	$0.5190 \to [0.5152, 0.5247]$	$0.4931 \to [0.4656, 0.4807]$	$0.0259 \to [0.0419, 0.0496]$

Table 14: Gain of the SEGO-returned hierarchy over the shape-matched random ensemble across 20 complete $k = 3$, $R = 64$ representation pipelines. Ranges and counts describe realized gains; 95% intervals resample pipelines and OOD identities. Official scaffold splits are fixed.

Detector	Mean (SD)	Realized range	Positive	Crossed 95% CI
Ledoit–Wolf	$+0.0264\,(0.0106)$	$[+0.0051, +0.0417]$	$20/20$	$[+0.0158, +0.0373]$
kNN-5	$+0.0158\,(0.0102)$	$[-0.0032, +0.0387]$	$18/20$	$[+0.0049, +0.0271]$
Isolation Forest	$+0.0274\,(0.0123)$	$[+0.0080, +0.0529]$	$20/20$	$[+0.0154, +0.0394]$

Reproducibility details. The SEGO implementation is pinned to commit `45bdda7cd36c` (Hou et al., 2025b). The subsequent analysis used Python 3.12.2, NumPy 1.26.4, SciPy 1.13.1, and scikit-learn 1.6.1; realization zero reproduces the primary $R = 64$, $k = 3$ point estimates exactly. Matching seeds are 17, 271, 911, 2027, and 8081; Isolation Forest uses 200 trees and seed zero.

For graph i at height k, primary randomization uses seed $70{,}000{,}000 + 1{,}000{,}000k + i$. Independent batch b adds $100{,}000{,}000b$. Pipeline realization $t > 0$ uses the same randomization stride, relabels with seed $91{,}000{,}000 + i + 1{,}000{,}000t$, and adds $1{,}000{,}003t$ to each matching seed. Its crossed bootstrap uses 5,000 draws and seed 91,417, incremented by 100 per detector.